DOCUMENT RESUME

ED 333 013 TM 016 461

AUTHOR Ackerman, Terry A.

TITLE An Examination of the Effect of Multidimensionality on Parallel Forms Construction.

PUB DATE Apr 91

NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1991).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Algorithms; Equations (Mathematics); Estimation (Mathematics); *Item Response Theory; *Mathematical Models; *Scores; Simulation; *Test Construction; Test Items

IDENTIFIERS Ability Estimates; Information Function (Tests); Item Parameters; *Multidimensionality (Tests); *Parallel Test Forms; Unidimensionality (Tests)

ABSTRACT

This paper examines the effect of using unidimensional item response theory (IRT) item parameter estimates of multidimensional items to create weakly parallel test forms using target information curves. To date, all computer-based algorithms that have been devised to create parallel test forms assume that the items are unidimensional. This paper focuses on one such algorithm, which was developed by R. M. Luecht and T. M. Hirsch. Unidimensional item parameter estimates were obtained by calibrating response data generated from two-dimensional item parameters. Using these unidimensional estimates, three sets of test items from a pool of 200 multidimensional items were selected for each of two different test lengths for three differently shaped target information functions. The item parameter estimates were obtained by calibrating five forms of the EAAP Mathematics usage test using the multidimensional IRT calibration program NOHARM. Response data were generated for 2,000 abilities. Observed score differences for each triad, based on the multidimensional item parameters, were then compared. Despite the multidimensionality of the selected items, the created forms appear to be quite parallel both unidimensionally and multidimensionally. Two tables and seven figures are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED333013

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TERRY A. ACKERMAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

An Examination of the Effect of Multidimensionality on Parallel Forms Construction

Terry A. Ackerman

University of Illinois

BEST COPY AVAILABLE

Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1991

TM016461

Abstract

This study examines the effect of using unidimensional IRT item parameter estimates of multidimensional items to create weakly parallel test forms using target information curves. To date all of the computer based algorithms that have been devised to create parallel test forms assume that the items are unidimensional. This study focuses on one such algorithm developed by Luecht and Hirsch. Unidimensional item parameter estimates were obtained by calibrating response data generated from two-dimensional item parameters. Using these unidimensional estimates three sets of test items were selected for each of two different test lengths for three different shaped target information functions. Observed score differences for each triad, based upon the multidimensional item parameters, were then compared. Results suggest that despite the multidimensionality of the selected items the created forms appear to be quite parallel both unidimensionally and multidimensionally.

An Examination of the Effect of Multidimensionality on Parallel Forms Construction

Measurement practitioners are beginning to realize the advantage afforded test construction techniques by using the computer. For the past five to six years several researchers have proposed several approaches that take advantage of the computational power of the computer to construct parallel test forms within an item response theory (IRT) framework (Theunissen, 1985; Ackerman, 1989; van der Linden & Boekkooi-Timminga, 1989; Luecht & Hirsch, in press; Adema, 1990). Concurrently, research examining the effects of multidimensionality has been suggesting that practitioners need to consider the consequences of constructing tests using multidimensional items. One such example is the area of item bias. Several researchers have demonstrated that item bias is the result of multidimensional items interacting with groups which have different underlying ability distributions (Ackerman, 1988; Shealy & Stout, 1989).

This paper extends the examination of multidimensionality to parallel forms construction. Currently all of the parallel forms construction approaches work with unidimensional IRT models. The assumption is made that all of the items within the pool are measuring the same trait or the same composite of multiple traits. Each item is considered for selection based upon its estimated unidimensional item parameters. As to date no approach selects items based upon model fit, its tacitly assumed that all items fit the unidimensional IRT model equally well. Thus, even if the selection pool contained multidimensional items only their unidimensional estimates would be considered in the test construction process.

Items which measure cognitive skills tend to be multidimensional (cf. Traub, 1983). One could even claim that no two items are ever exactly unidimensional. That is, in a multidimensional IRT sense, no two items will ever be measuring exactly the same composite of skills. The issue of tests having multidimensional items raises some important questions that practitioners need to be examine. For example, how much multidimensionality can a test have before the test possesses the potential for bias? Can multidimensionality cause the unidimensional score scale to have a different meaning for low scores than for high scores? How do multidimensional

items interfere with the unidimensional equating of test forms? What happens when tests are constructed from unidimensional estimates of multidimensional items?

The last question provides the central theme of this paper. Parallel forms methodology insures that test forms created by matching an IRT target information curve will be parallel. But how parallel will the test forms be if the multidimensional nature of the items which compose each form is considered? The purpose of this paper is to investigate observed score differences in test forms that are created to be parallel using unidimensional estimates of multidimensional items.

Background

In item response theory, tests are defined as being parallel if they have similar test information functions. Information, within the IRT context, is conditional upon the level of ability, θ, and represents the degree of measurement precision of an item or a collection of items at the given θ level. Formally IRT item information is defined as

$$I_j(\theta_k) = \frac{P_j'(\theta_k)^2}{P_j(\theta_k)Q_j(\theta_k)} \qquad (1)$$

where $P_j(\theta_k)$ is the probability of a correct response to item "j" at the given ability level, θ_k, $Q_j(\theta_k) = 1 - P_j(\theta_k)$, and $P_j'(\theta_k)$ is the first derivative of $P_j(\theta_k)$ with respect to θ_k. Item information functions, $I_j(\theta_k)$ are additive. Thus the information for a given n-item test, the test information function can be expressed as

$$T(\theta_k) = \sum_{j=1}^{n} I_j(\theta_k) \qquad (2)$$

Tests that have identical test information functions and are measuring the same skills (i.e., same content) may be considered weakly parallel (Samejima, 1977). Within this context,

same content) may be considered weakly parallel (Samejima, 1977). Within this context, researchers have developed algorithms to create test forms which have the same test information function, and thus can be considered to be weakly parallel.

Some of the proposed algorithms employ zero-one linear programming techniques to maximize the test information in the created forms (Theunissen, 1985; van der Linden and Boekooi-Timminga, 1989). These techniques tend to require large amounts of computing time and do not seem applicable large scale testing programs such as the American College Testing Program (ACT) or Educational Testing Service (ETS).

A more heuristic approach was suggested by Ackerman (1989). In this approach the item information for each test form was prioritized at different θ levels according to the distance between the current test information values and a specified target information curve. Items in the pool were presorted at these θ levels by descending amount of information. Items which contributed the most information at the designated priority points on the test information curve were assigned to the test forms. One draw back to this approach was that the target information curves tended to be "overshot" and the created forms would contain too much information.

Luecht and Hirsch (in press) suggested a third approach in which items are selected to fill a target curve based upon the best uniform growth throughout a specified ability range. Such an algorithm prevented sporadic information growth in an one region of the ability scale encountered in the Ackerman (1989) heuristic. Because of the success of this last algorithm to match specified target curves it was the algorithm employed in this study.

Method

To study the effect of multidimensionality of parallel forms construction a pool of 200 multidimensional items was formed. These item parameter estimates were based upon the two-dimensional compensatory IRT model. They were obtained by calibrating five forms of the EAAP Mathematics usage test using the multidimensional IRT calibration program NOHARM (Fraser, 1983). The two-dimensional compensatory IRT model defines the probability of a

correct response to item i by examinee j as

$$P(X_{ij}=1|a_i,d_i,\boldsymbol{\theta}_j)=\frac{e^{(a_i'\boldsymbol{\theta}_j+d_i)}}{1+e^{(a_i'\boldsymbol{\theta}_j+d_i)}} \quad (2)$$

where X_{ij} is the score (0,1) on item i by person j,

$\mathbf{a}_i$ is the vector of item discrimination parameters,

d_i is a scalar difficulty parameter for item i, and

$\boldsymbol{\theta}_j$ is the vector of ability parameters for person j.

One common method used to visually assess the dimensionality of the pool is to plot the two-dimensional vectors for each item (Reckase, 1985). The vectors representing the 200 items used is this study are shown in Figure 1. The length of the item vector indicates the degree of multidimensional discrimination. The angle the vector makes with the positive θ_1 axis indicates the composite of θ_1-θ_2 being measured by the item. The base of each vector is orthogonal to the p = .5 equiprobability contour of the item response surface. If the test were strictly unidimensional all of the vectors would tend to be aligned in exactly the same direction. As can be seen in Figure 1, the pool was clearly two-dimensional.

Insert Figure 1 about here

The two-dimensional item parameter estimates were converted to unidimensional estimates by calibrating generated multidimensional response data to fit a unidimensional 2PL IRT model. Specifically, response data was generated for 2000 (θ_1,θ_2) abilities, randomly selected from a bivariate normal distribution centered at (0,0) and having an identity variance-covariance matrix. This response data was then fit to the 2PL model using the calibration program BILOG (Mislevy & Bock, 1983). The estimated unidimensional item parameters were then used to

create parallel tests forms using the algorithm of Luecht and Hirsch (in press).

Three parallel test forms were created to fit each of six specified target information curves (three different shapes X two different test lengths) using the program ITEMSEL (Luecht & Hirsch, in press). The three selected shapes included a negatively skewed test information function, a normal shaped test information function and a positively skewed test information function. The shapes were intended to simulate the expected target shapes for three different testing situations: an admissions test (positively skewed), a standard achievement test (normal), and a scholarship test (negatively skewed). Three forms were created for each test shape for two different test lengths, 25 items and 40 items. The three different shaped target curves for the for the 25-item tests along with the total pool information function are shown in Figure 2.

Insert Figure 2 about here

Once each triad of tests was created for each shape and each test length, the two-dimensional response data that was previously generated for the unidimensional calibration runs was used as the response data for each created form. Summary statistics on the observed score distributions were then computed. Likewise several multidimensional analyses were conducted.

Results

The first four moments, the KR-20 reliability estimate, the minimum and maximum angle of the selected items and the reference composite angle are shown for each of the three 25-item tests for each target shape in Table 1. The results for the 40-item tests are shown in Table 2.

Insert Tables 1 & 2 about here

The largest mean difference was .61 for the 25-item tests (between Test 1 and Test 3 for

the positively skewed case) and .79 for the 40-item tests (between Test 2 and Test 3 for the negatively skewed target). Values for the first four moments of each observed score generated distribution suggest a high degree of similarity in shape of each simulated observed score distribution. KR-20 reliability estimates are likewise quite similar with the largest difference being .03 in any one triad. The three 25-item tests created to match the positively skewed target had the lowest reliability average (.80); whereas, the three 40-item tests created for the negatively skewed target had the highest reliability average (.90).

As might be expected the 40-item tests had more of an angular spread of item vectors than the 25-item tests. The largest spread of vectors was 87.48 degrees for both Test 1 in the negatively skewed condition and test 3 in the normal condition. The largest angular spread for the 25-item tests was 61.21 degrees for Test 2 with the normal curve shaped target. The narrowest 25-item spread was 26 degrees for Test 1 for the positively skewed target. It appears that the shape of the target curve had little effect on the homogeneity of the angular composites for each triad.

The reference composite (Wang, 1986) is an important characteristic used to assess the effect of the multidimensionality. Wang determined analytically how the two-dimensional latent ability space would be mapped onto the unidimensional ability scale. The angle associated with the reference composite defines the unidimensional score scale in terms of a θ_1-θ_2 composite. Thus, if the angle of the θ_1-θ_2 reference composite is 45 degrees the unidimensional score scale could be interpreted as an equal weighting of the skills defined by θ_1 and θ_2. If the forms are truly parallel each would be measuring the same composite of θ_1 and θ_2. Amazingly, the angle associated with the reference composite is quite similar in each triad across of shapes and for each size test. The pool reference composite was 38.38 degrees, which was quite close to the reference composite for most of the created forms. The largest difference for both the 25- and the 40-item test was between Test 1 and Test 3 for the 25-item positively skewed target, 9 degrees.

Because the measurement direction of the reference composite is influenced by the spread of the item vectors, vector plots were created for each of the forms to see how multidimensionally heterogeneous the items actually were. The item vector plots for Test 1 and Test 3 (of the 25-item positively skewed condition) are displayed in Figures 3 and 4. Test 3 had

the larger angular spread and appeared to contain more discriminating items. The vectors from both tests however, appeared to lie in a relatively similar sector of the two-dimensional latent ability space.

Insert Figures 3 & 4 about here

To further investigate differences between created forms two additional multidimensional graphical analysis were conducted. The first involved computing the (θ_1,θ_2) conditional distributions for each possible raw score category using the two-dimensional item parameters for the selected items and a recursive formulation suggested by Stocking and Lord (1983). The conditional means and conditional θ_1 and θ_2 variances were also determined. The plot of the (θ_1,θ_2) centroids for each of the raw score categories 1 - 25 for the two tests having the largest angular difference in their reference composite (Test 1 and Test 3 for the positively skewed 25-item target) are plotted in Figures 5 and 6.

Around every fifth centroid is an ellipse indicating the size of the conditional θ_1 and θ_2 variances. The length of the horizontal axis of the ellipse represents the size of the conditional θ_1 variance and the length of the vertical axis represents the size of the conditional θ_2 variance.

Insert Figures 5 & 6 about here

Even though the angles of the reference composites might suggest that these two forms would measure slightly different θ_1-θ_2 skills, centroids of the conditional distributions for the respective raw score categories appear to be located in relative close proximity to one another. This implies that examinees located in similar regions of the two-dimensional ability plane would be mapped into the same raw score categories.

The computed conditional variances also appear to be close in value suggesting that the

observed scale represents a consistent interpretation throughout the observed score range for each test. Specifically, θ_1 and θ_2 seem to be measured with the same degree of precision at each possible observed score. In both forms the vertical axes of the variance ellipses are slightly longer than the horizontal axes, suggesting that θ_2 is measured more accurately than θ_1. This analysis was conducted on several triads with similar results.

One final multidimensional analysis that was conducted was to examine the difference between the true score surface for different pairs of parallel forms. If the test forms are truly parallel, the differences between the true score surfaces for each form should be relatively minor. The surface representing such a difference is plotted in Figure 7 for the Test 1 and 3 having the 25-item positively skewed target information curve.

Insert Figure 7 about here

If the two true score surfaces where identical the surface representing the difference would lie in the outlined "zero" plane. Because the difference surface was the true score surface of Test 1 minus the true score surface of Test 3 regions of the two-dimensional ability plane which lie above the zero plane represent (θ_1,θ_2) ability regions in which examinees would perform better on Test 1. The reverse is true where the surface dips below the zero plane. From the corresponding contour plot, it appears that examinees in the first and fourth quadrants of the ability plane would have a true score of 2 to 10 points higher on Test 3. Examinees in the third and fourth quadrants would have true score values 10 points on Test 1. Although these differences appear to be quite large the standard error of the true score difference estimated using the reliabilities for both tests would be about 6 true score points. Thus one would expect that 95% of the examinees would lie within + or - 12 true score points if there was no difference in the true score surfaces.

Discussion

The purpose was this paper was to provide practitioners with insight about the effect of

using unidimensional item parameter estimates of multidimensional items for constructing parallel forms using the IRT information function. Amazingly the multidimensionality of the pool does not play havoc with the parallel forms construction process, at least using the algorithm suggested by Luecht and Hirsch (1989). As the test length increased from 25 to 40 items the spread of θ_1-θ_2 composite being measured by the selected items appeared to increase. However, the reference composite appeared to be aligned in similar direction in the two-dimensional ability plane for most of the created triads.

Differences were also examined from a multidimensional perspective, such as plotting the centroids and variance ellipses of the conditional distributions for each raw score category. These analyses also failed to reveal significant differences.

One possible explanation may be in the way the Leucht and Hirsch algorithm selects items. Upon examining the order of item selection for each test it appeared that the procedure tends to select items at or near the pool reference composite first. As subsequent items are selected, the angle of the selected items start to move further and further away from the reference composite. However, items appear to be selected in a balanced fashion, with each form being assigned about equal number of items on each side of the reference composite. Such a process seems to insure that the reference composite for each form will be roughly pointed in the same direction in the two-dimensional latent space.

Future research needs to investigate the possibility of restraining the item selection process to defined sectors within the two-dimensional ability plane. That is, items would only be considered if their two-dimensional item vectors were within a specified number of degrees of a predetermined reference composite. Such a restriction would decrease the spread of items selected for any one form, but should increase the internal consistency of each form.

References

Ackerman, T. A. (1989, April). An alternative methodology for creating parallel test forms using the IRT information function. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Ackerman, T.A. (1988, April) An explanation of differential item functioning from a multidimensional perspective. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Adema, J.J. (1990). A revised simplex method for test construction problems. (Research Report No. 90-5) Enschede, the Netherlands: University of Twente, Department of Education.

Fraser, C. (1983). NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: University of New England, Centre for Behavioral Studies.

Luecht, R. M. & Hirsch, T.M. (in press). Item selection: a proposed heuristic for meeting target information points on a curve. Applied Psychological Measurement.

Mislevy, R.J. & Bock, R.D. (1983). BILOG: Item analysis and test scoring with binary logistic models. [Computer program]. Mooresville, IN: Scientific Software.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. Psychometrika, 42, 193-198.

Shealy, R. & Stout, W. (1989, April). A procedure to detect test bias present simultaneously in several items. Paper presented at the annual meeting of the American Educational Research Association: San Francisco.

Stocking, M. & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Theunissen, T.J.J.M.(1985). Binary programming and test design. Psychometrika, 50(4), 411-420.

Traub, R.E. (1983). A priori consideration in choosing an item response model. In R.K. Hambleton (ED.). Applications of item response theory (pp. 57-70). Vancover BC: Educational Research Institute of British Columbia.

van der Linden, W.J. and Boekkooi-Timminga, E. (1989). A maxmin model for test design with practical constraints. Psychometrika, 54(2), 237-247.

Wang, M. (1986, April). Fitting a unidimensional model to multidimensional item response data. Paper presented at the ONR Contractors Conference. Gatlinburg, TN.

Table 1
Summary statistics of the generated observed score distributions for the three 25-item target information shapes.

	Shape of Target Information Curve								
	Positively Skewed			Normal			Negatively Skewed		
Test #	1	2	3	1	2	3	1	2	3
$\bar{X}$	11.89	11.37	11.28	11.04	10.72	10.99	10.75	10.33	10.69
s	5.07	4.98	4.98	5.91	5.99	5.75	5.95	6.31	5.90
g_1	.12	.10	.14	.25	.38	.26	.29	.37	.36
g_2	-.71	-.69	-.68	-.86	-.75	-.85	-.83	-.84	-.81
KR-20	.80	.79	.79	.86	.87	.85	.86	.89	.86
Min α	14.68	0.00	0.00	4.41	0.00	4.46	12.75	4.41	0.00
Max α	46.58	57.66	61.21	57.66	61.21	64.78	71.51	61.21	57.66
RC α	31.98	39.80	41.46	39.33	40.46	40.27	42.00	38.23	38.93

Note: g_1 = skewness; g_2 = kurtosis; RC α = reference composite angle; N=2000

Table 2
Summary statistics of the generated observed score distributions for the three 40-item target information shapes.

	Shape of Target Information Curve								
	Positively Skewed			Normal			Negatively Skewed		
Test #	1	2	3	1	2	3	1	2	3
$\bar{X}$	16.93	16.65	17.13	16.02	15.57	15.47	16.97	16.15	16.94
s	8.11	8.17	7.83	8.28	8.65	8.42	8.85	9.06	8.69
g_1	.28	.35	.23	.41	.46	.43	.34	.37	.30
g_2	-.71	-.65	-.66	-.59	-.59	-.63	-.78	-.79	-.78
KR-20	.89	.88	.87	.89	.91	.90	.90	.91	.90
Min α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max α	87.48	59.32	64.78	74.48	64.78	87.48	59.32	74.48	71.51
RC α	42.87	34.81	36.20	40.16	36.48	39.37	37.78	42.64	39.36

Note: g_1 = skewness; g_2 = kurtosis; RC α = reference composite angle; N=2000

Figure Captions

Figure 1. Two-dimensional item vectors for the 200-item pool.

Figure 2. The unidimensional test pool information function and the three specified 25-item target information curves.

Figure 3. Two-dimensional item vectors for the 25 items for Test 1 selected to match the positively skewed target information function.

Figure 4. Two-dimensional item vectors for the 25 items for Test 3 selected to match the positively skewed target information function.

Figure 5. The (θ_1,θ_2) conditional centroids and variance ellipses for the 25 observed score categories for Test 1 in the positively skewed information condition.

Figure 6. The (θ_1,θ_2) conditional centroids and variance ellipses for the 25 observed score categories for Test 3 in the positively skewed information condition.

Figure 7. The true score difference surface (Test 1 - Test 3) and corresponding contour plot.

Figure 1

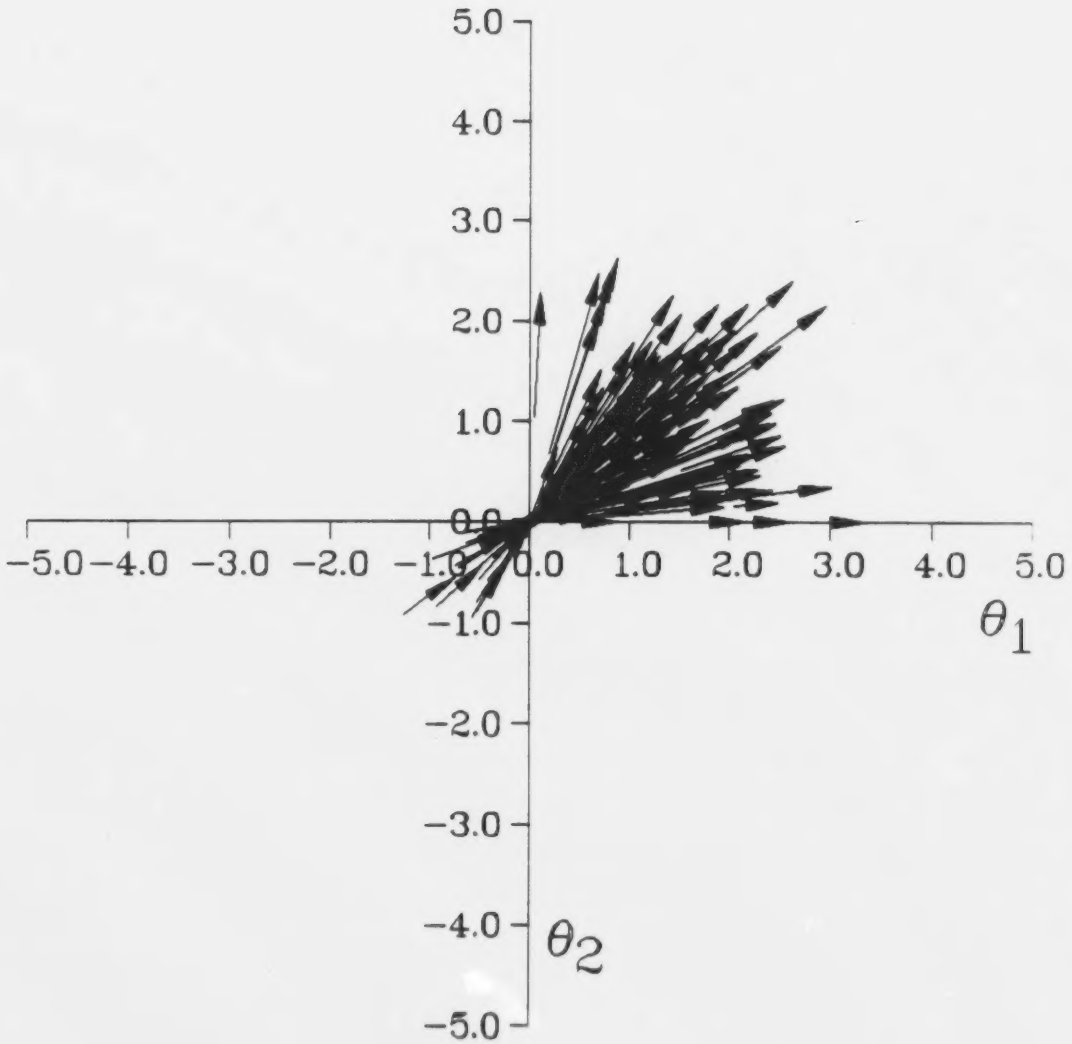
5.0
4.0
3.0
2.0
1.0
0.0
−1.0
−2.0
−3.0
−4.0
−5.0
−5.0 −4.0 −3.0 −2.0 −1.0 0.0 1.0 2.0 3.0 4.0 5.0
θ_1
θ_2

Figure 2

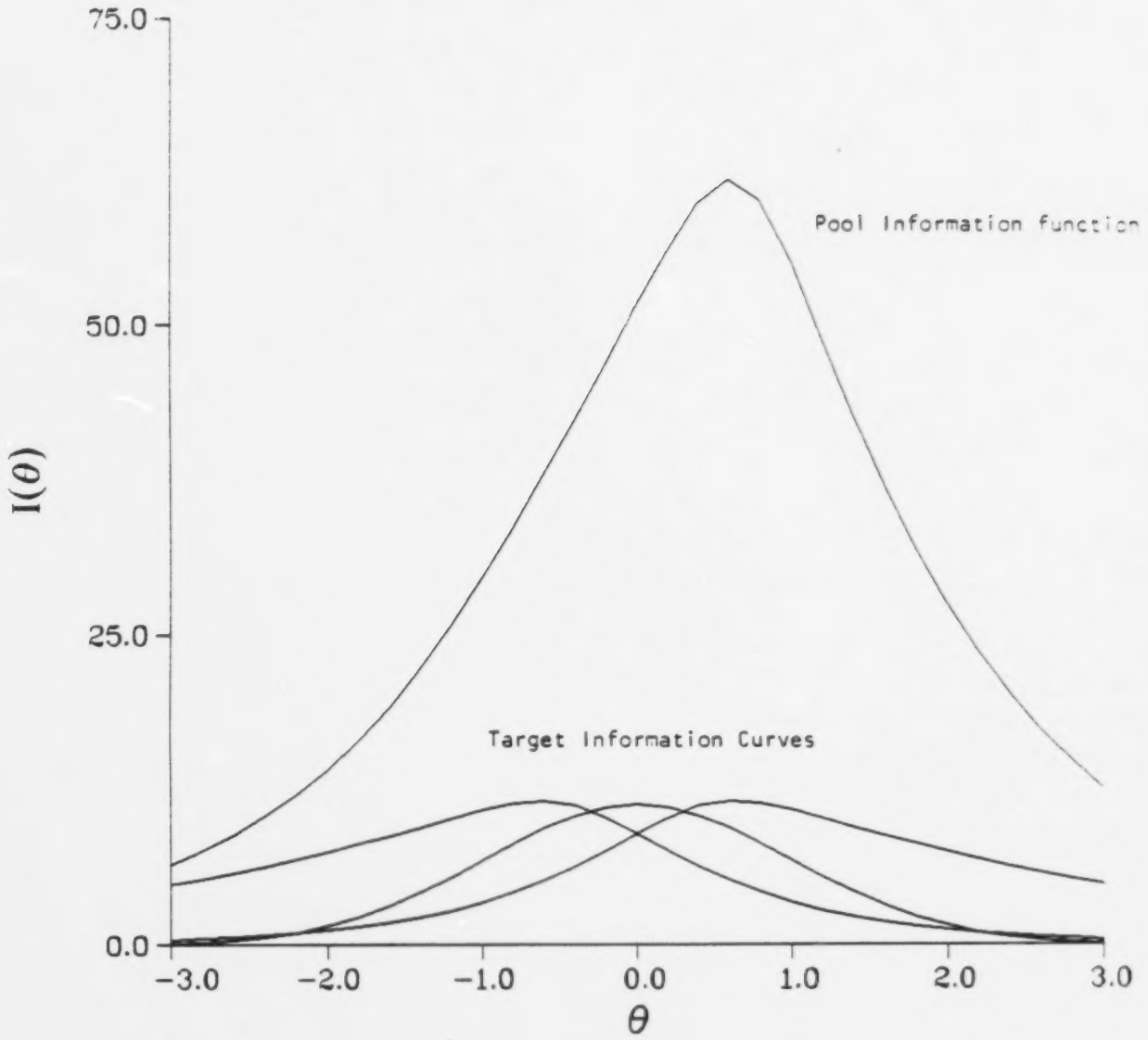
75.0
50.0
25.0
0.0
I(θ)
Pool Information function
Target Information Curves
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 3

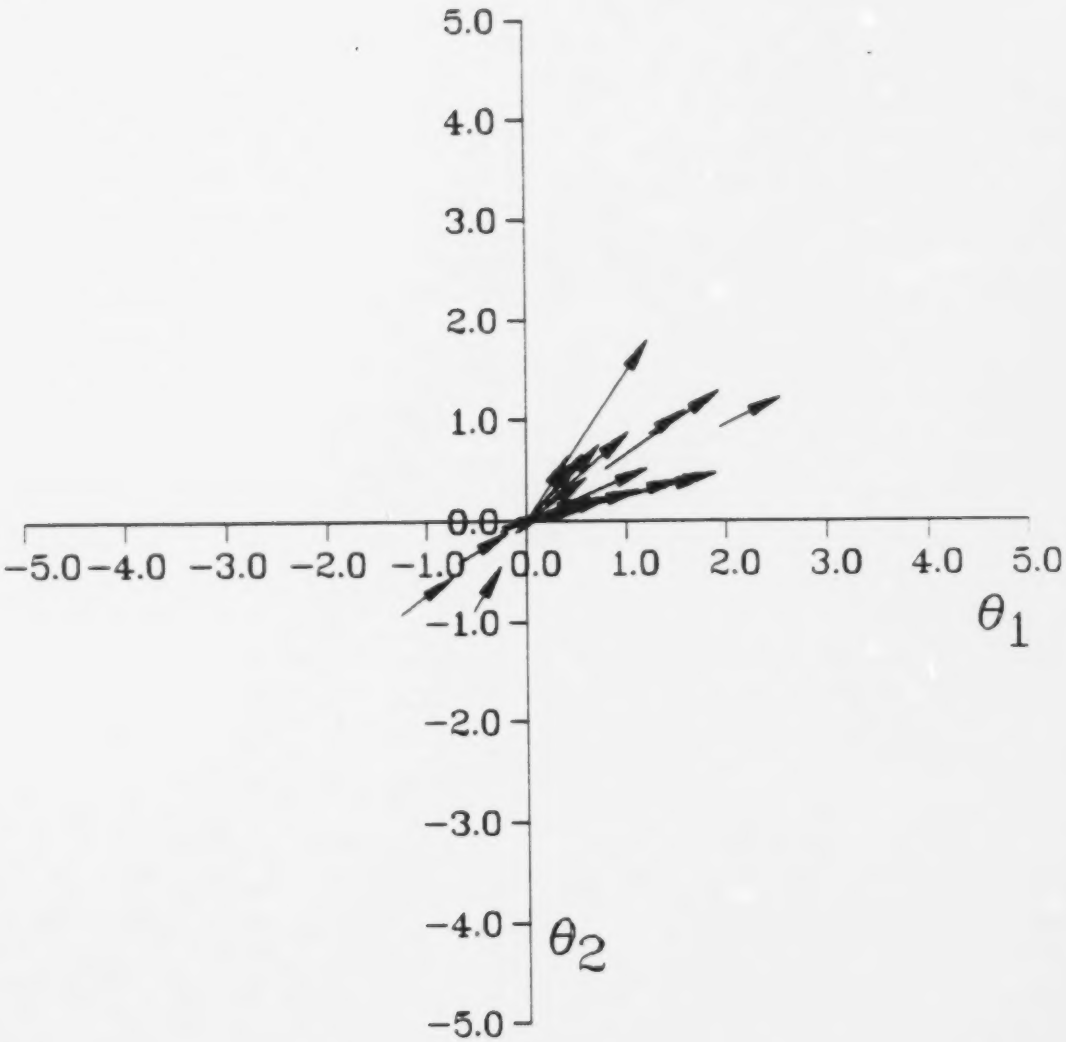
5.0
4.0
3.0
2.0
1.0
0.0
−1.0
−2.0
−3.0
−4.0
−5.0
−5.0 −4.0 −3.0 −2.0 −1.0 0.0 1.0 2.0 3.0 4.0 5.0
θ_1
θ_2

Figure 4

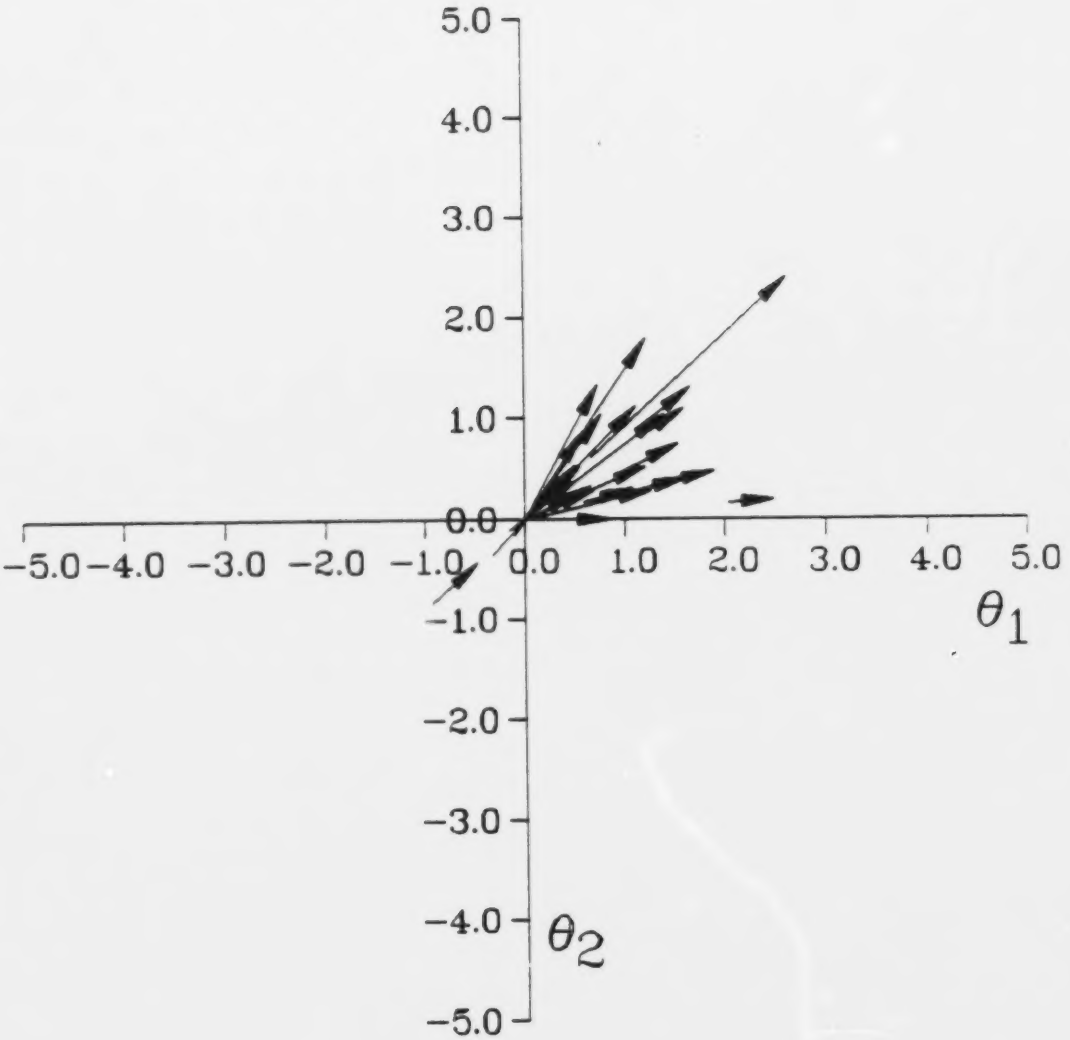
5.0
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-5.0 -4.0 -3.0 -2.0 -1.0 0.0 1.0 2.0 3.0 4.0 5.0
θ_1
θ_2

Figure 5

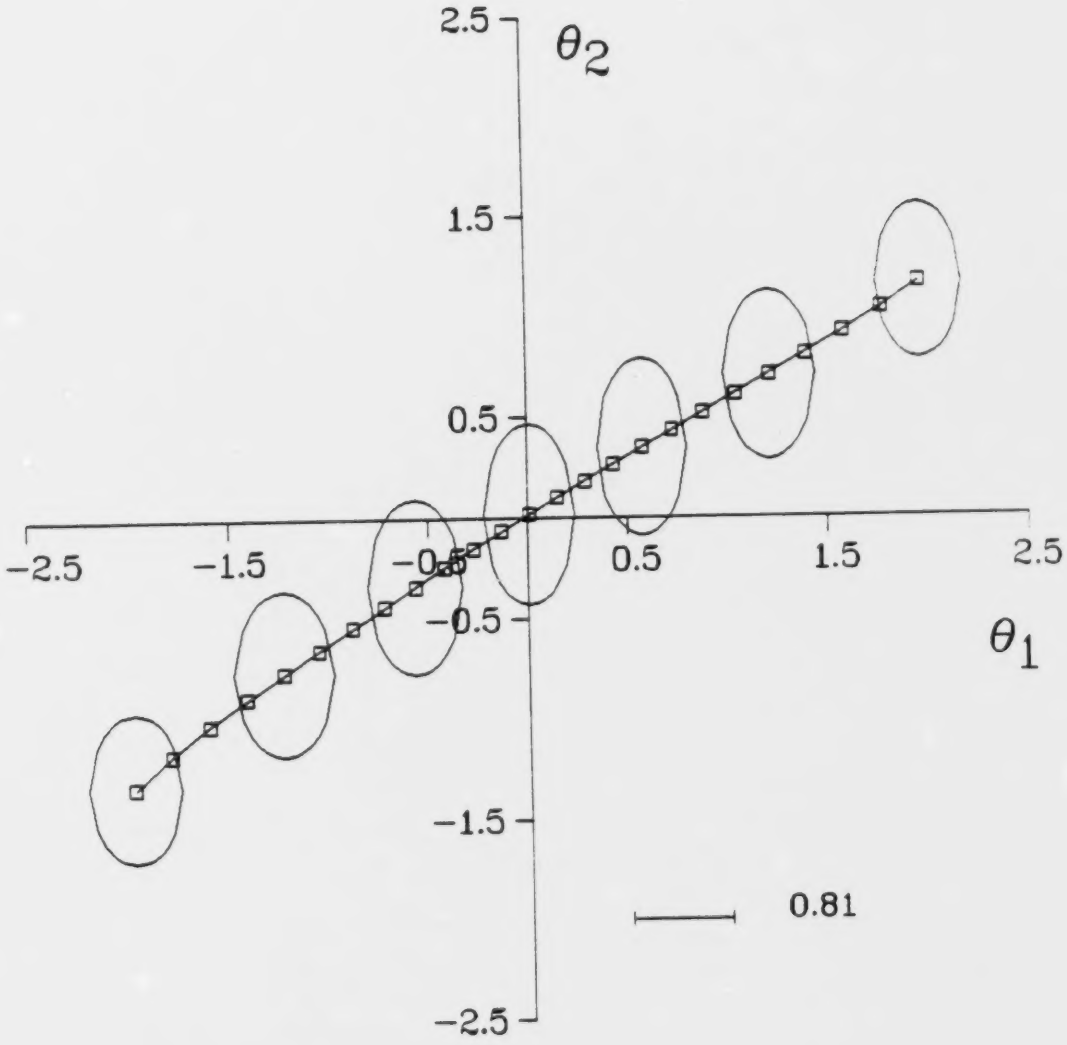

Figure 6

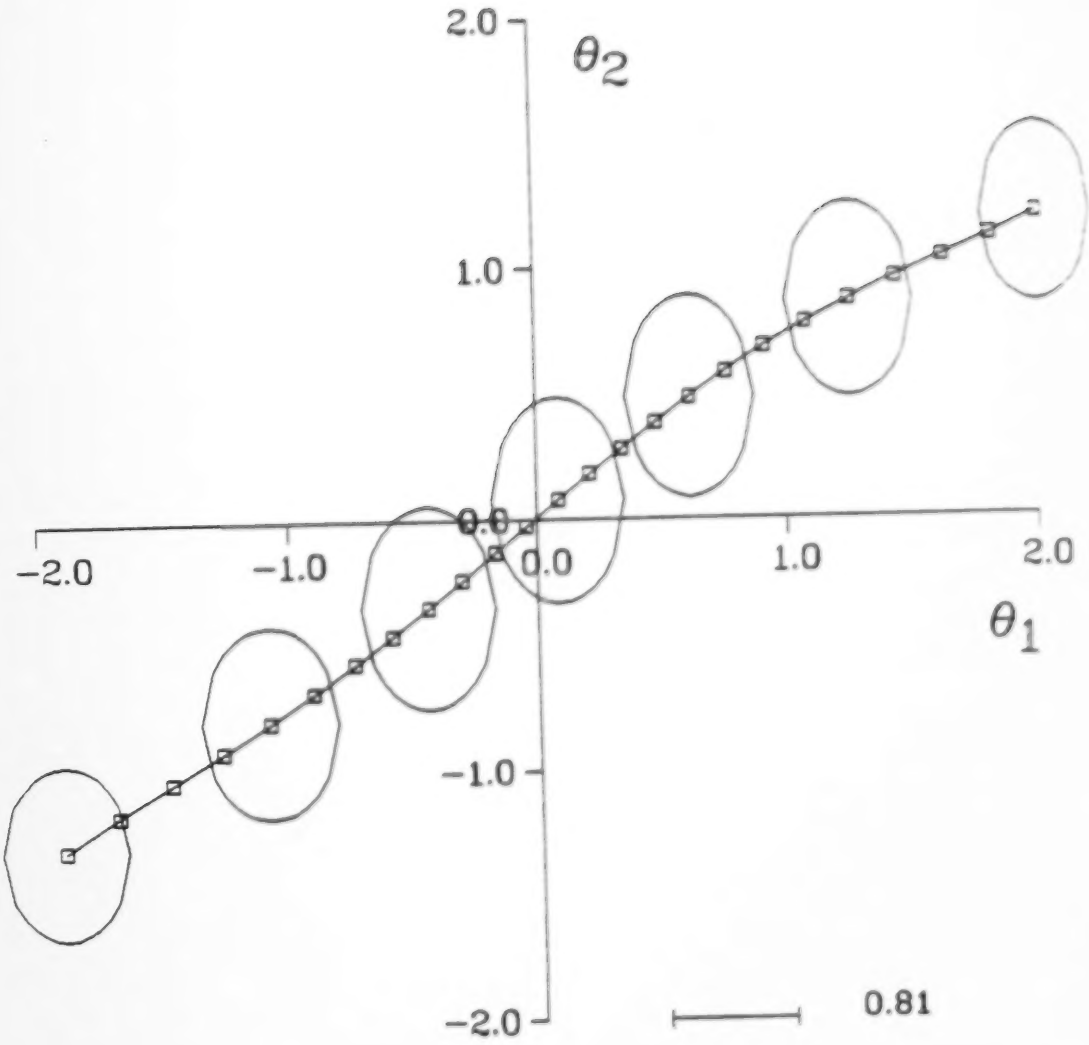

Figure 7